Advanced Regression: A note on collinearity

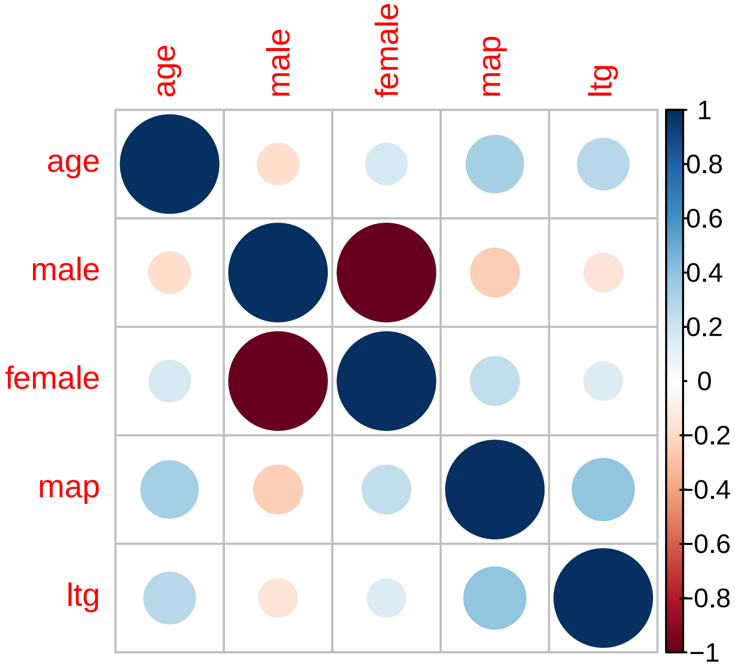
Garyfallos Konstantinoudis

Mar 5, 2024

Introduction

We consider again the diabetes outcome looking at the outcome disease progression y and we try to fit the following linear model:

$y = \alpha + age + male + female + map + ltg$



Fitting an lm when high correlation

```
lm(y ~ age + male + female + map + ltg, data = x) %>% summary()
```

```

Call:
lm(formula = y ~ age + male + female + map + ltg, data = x)

Residuals:
     Min       1Q   Median       3Q      Max 
-166.017  -42.787   -5.523   41.751  185.752 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  144.502      4.303  33.586  < 2e-16 ***
age          -31.454     65.564  -0.480   0.6317    
male          14.353      6.000   2.392   0.0172 *  
female            NA         NA      NA       NA    
map          460.104     69.384   6.631 9.84e-11 ***
ltg          766.189     66.966  11.442  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.7 on 437 degrees of freedom
Multiple R-squared:  0.3856,	Adjusted R-squared:   0.38 
F-statistic: 68.57 on 4 and 437 DF,  p-value: < 2.2e-16
```

- Option in `lm()` function: `singular.ok = TRUE` automatically removes female.

```
lm(y ~ age + male + female + map + ltg, data = x, singular.ok = FALSE)
```

```
Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...): singular fit encounte
```

Fitting an lm when high correlation

- The `lm()` function checks for singularities in the design matrix x, but not all methods have this safety check.

- Example: Lasso regression

```r
library(glmnet)
lm.lasso <- glmnet(y = y, x = x, alpha = 1, lambda = 0.5, family = "gaussian")
lm.lasso$beta
```

```
12 x 1 sparse Matrix of class "dgCMatrix"
                  s0
age      .
sex    -1.230792e+00
bmi     5.250119e+02
map     3.083265e+02
tc     -1.609111e+02
ldl      .
hdl    -1.805392e+02
tch     6.606582e+01
ltg     5.242775e+02
glu     6.113052e+01
male    2.050998e+01
female -2.153821e-12
```

What are singularity and multicollinearity?

Singularity

One predictor variable in a multiple regression model can be exactly explained by the other $p-1$ predictor variables.

Multicollinearity

One predictor variable in a multiple regression model can be linearly explained by the other $p-1$ predictor variables with high accuracy.

What can cause singularity?

- Dummy-coding of categorical variable. Make sure not to add redundant information
- Do not include multiple measurement that are measured on different scales (e.g., m and cm)

What is the impact of multicollinearity?

True biological processes do not cause singularity (because they are random, not deterministic), but can cause multicollinearity.

- The computation of the ordinary least squares estimate requires an inversion of the $p \times p$-dimensional correlation matrix $x^t x$.
- $x^t x$ cannot be inverted when $x^t x$ is singular.
- When there is multicollinearity, $x^t x$ can be inverted, but the estimate will show a high variance and will be highly unstable.
- Multicollinearity can distort a linear model and impact the interpretation.

How to inspect correlation?

For generic correlation structures:

- Correlation and covariance matrix

How to detect singularity?

- Rank of matrix

How to detect multicollinearity

- Variance inflation factor

Covariance matrix

Computing the sample covariance matrix using matrix multiplication

$$\hat{cov}(x) = \frac{1}{n-1} \underbrace{x_c^t}_{p\times n} \underbrace{x_c}_{n\times p}$$

- x_c is centred (mean is zero) $x_c = x - 1_n \bar{x} = cx$
  - where $\bar{x} = (\bar{x}_1, ..., \bar{x}_p)$ is the vector of means
  - and 1_n is a vector of ones
  - and $c = I_n - \frac{1}{n} 1_n 1_n^t$
  - and I_n is the $n \times n$ identity matrix with ones on the diagonal
- x predictor matrix of n rows and p columns
- x^t transposed predictor matrix of p rows and n columns

Matrix multiplication

Matrix multiplication: $\underbrace{c}_{n\times p} = \underbrace{a}_{n\times m}\underbrace{b}_{m\times p}$

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$$

- a is a $n \times m$ and b is a $m \times p$ matrix
- c is a $n \times m \times m \times p = n \times p$ matrix
- Make sure your matrices have the correct dimensions, number of columns of the left matrix must be equal to the number of rows on the right.
- Can be computed in R using the `\%*\%` command.

Correlation matrix

Computing the sample correlation matrix using matrix multiplication

$$\hat{cor}(x) = \frac{1}{n-1} \underbrace{x_s^t}_{p\times n} \underbrace{x_s}_{n\times p}$$

where x_s is a centred and scaled matrix $x_s = cxd^{-1}$

- where $d=diag(s)$ is a diagonal matrix
- with the sample standard deviation s on the diagonal.

This is equivalent to writing

$$\hat{cor}(x_j, x_k) = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

Correlation matrix

- Correlation matrices are symmetric and have a vector of 1's on the diagonal.

```
cor(x %>% dplyr::select(age, male, female, map, ltg))
```

```
              age       male     female        map        ltg
age     1.0000000 -0.1737371  0.1737371  0.3354267  0.2707768
male   -0.1737371  1.0000000 -1.0000000 -0.2410132 -0.1499176
female  0.1737371 -1.0000000  1.0000000  0.2410132  0.1499176
map     0.3354267 -0.2410132  0.2410132  1.0000000  0.3934781
ltg     0.2707768 -0.1499176  0.1499176  0.3934781  1.0000000
```

- Note the following correlation matrix captures the correlation between the samples and is of dimension $n \times n$

$$c\hat{o}r(x^t) = \frac{1}{p-1} \underbrace{x_s}_{n\times p} \underbrace{x_s^t}_{p\times n}$$

Correlation matrix

R commands

- `cov()` sample covariance matrix
- `cor()` sample correlation matrix
- `corrplot()` to visualise

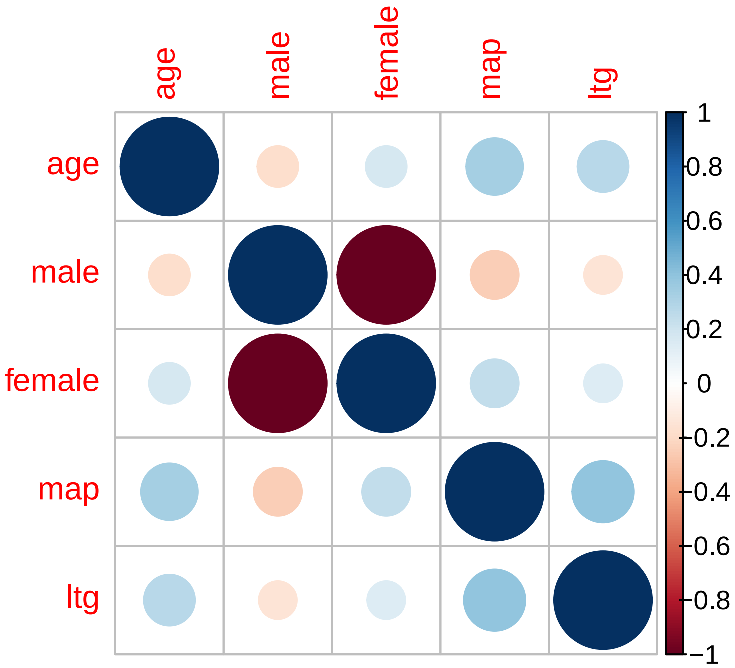
The rank of a matrix

- Consider a matrix x of dimension $n \times p$.

$$\underbrace{x}_{n\times p}$$

- The rank of matrix x is the minimum of n and p.
- If we have more samples than variables $(n > p)$ the rank is p.

If we have less samples than variables $(n < p)$ the rank is n.



The rank of the correlation matrix

- Let us consider again the correlation matrix

$$c\hat{o}r(x) = \frac{1}{n-1} \underbrace{x_s^t}_{p\times n} \underbrace{x_s}_{n\times p}$$

- The theoretical rank of the correlation matrix is the minimum of n and p.
- To test the rank of a matrix in R: `rankMatrix()` in the `Matrix` package

If the rank of a correlation matrix is smaller than $min(n, p)$ the correlation matrix is singular and thus cannot be inverted.

The rank of the correlation matrix in R

```
x %>%
  dplyr::select(age, male, female, map, ltg) %>%
  as.matrix() %>%
  cor()
```

```
              age       male     female        map        ltg
age     1.0000000 -0.1737371  0.1737371  0.3354267  0.2707768
male   -0.1737371  1.0000000 -1.0000000 -0.2410132 -0.1499176
female  0.1737371 -1.0000000  1.0000000  0.2410132  0.1499176
map     0.3354267 -0.2410132  0.2410132  1.0000000  0.3934781
ltg     0.2707768 -0.1499176  0.1499176  0.3934781  1.0000000
```

```
x %>%
  dplyr::select(age, male, female, map, ltg) %>%
  as.matrix() %>%
  cor() %>%
  rankMatrix()
```

```
[1] 4
attr(,"method")
[1] "tolNorm2"
attr(,"useGrad")
[1] FALSE
attr(,"tol")
[1] 1.110223e-15
```

Interpretation: The correlation matrix of the design matrix with 5 predictors is of dimension 5×5, yet the rank is 4 which indicates singularity.

Variance inflation factor (VIF)

- The VIF is the ratio of the variance of β_j when fitting the full model divided by the variance of $\beta_{UNI}(j)$ in a unvariable linear model.
- Lowest possible value is 1 (no collinearity).
- Rule of thumb: If VIF > 10, this indicates strong multicollinearity, but already smaller VIF can impact the analysis.
- It provides an indication how much the variance of an estimated regression coefficient is increased because of multicollinearity.

Variance inflation factor (VIF)

Consider the following linear model including p predictors with inde $j \in 1, ..., p$

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_j x_j + ... + \beta_p x_p + \epsilon.$$

- For the first variable $j = 1$ fit a linear model, where x_1 is the outcome and all other variables x_{-1} are the predictors

$$x_1 = \alpha + \beta_2 x_2 + ... + \beta_j x_j + ... + \beta_p x_p + \epsilon.$$

- Estimate $R_2(1)$, the proportion of variance of x_1 explained by the other predictors x_{-1}.
- The VIF for variable 1 is defined as

$$VIF_1 = \frac{1}{1 - R_2(1)}$$

- Repeat for the other $j \in 2, ..., p$.

Variance inflation factor

R commands

- `vif()` in the R-package `car`
- Computes variance-inflation and generalized variance-inflation factors for linear and generalized linear models.

```r
library(car)
lm2 <- lm(y ~ age + male + map + ltg, data = x)
vif(lm2)
```

```
     age     male      map      ltg 
1.166584 1.075047 1.306446 1.216982 
```

- Interpretation: No variable has a VIF > 10, with around 1 they are rather low and there is no indication of multicollinearity.

Summary

- What are singularity and multicollinearity?
- How to detect singularity and multicollinearity?
  - Correlation and covariance matrix
  - The rank of a matrix
  - Variance inflation factor

Questions?